

Supporting Information

Kornev et al. 10.1073/pnas.0807988105

SI Text

Involvement Score Rationale. Involvement Score is a novel concept that is related to the nature of the LSP-alignment as a graph-theory based method. *Graph* is a way to represent information about a set of objects and relations between them. The objects are usually depicted as *vertices*, and relations between them as *edges* (Fig. S1). For example, vertices can represent airports and edges direct flights between them. Two identical graphs are called “isomorphic,” that means that they consist of identical objects and are connected in the same way. In other words there is a “one-to-one” correspondence between all elements of these graphs. Fig. S1 shows a pair of similar graphs that have such correspondence only between parts of the graphs, colored red. These identical parts are called “isomorphic subgraphs.”

In our method, protein molecule is presented as a graph. Residues are considered as vertices of the graph while edges describe spatial relations between them. By “spatial relation” we mean distance between these two residues and mutual orientation of their side chains. Orientation of a side chain in space can be defined by multiple ways. In our case, we define it as orientation of $C\alpha$ - $C\beta$ vector. The LSP alignment procedure is capable to compare two graphs that represent two different proteins and to discover isomorphic subgraphs between them. Thus we identify “one-to-one” correspondence between residues of these proteins, i.e., define an alignment between them. The unique feature of such alignment is that it is presented as a graph that consists not only from vertices but also from edges.

Each vertex can have different number of edges. Our experience shows that simple counting of the edges can provide important information about the residue. This number we define as the “Involvement Score.” As we mentioned earlier, edges describe spatial relations between residues. That is if two residues in the isomorphic subgraphs are connected, they are positioned in space in both proteins in the same way. If a residue has numerous connections on the isomorphic subgraph it means that the position of the residue in space and the orientation of its side chain is conserved with respect to multiple residues. This suggests that the residue is involved in formation of a precisely organized cluster of similar residues. From this we derive a conclusion that residues with high Involvement Scores can play an important functional and/or structural role.

The LSP-alignment method is still under development and the way of the Involvement Score calculation is the most simple: a sum of edges on isomorphic subgraph. Such approach does not consider precision of the geometrical fit between residues. As soon as they are positioned in space within a predefined tolerance, the edge is created. Another simplification is level of similarity between two residues. Instead of using elaborated similarity scores we considered them either similar or dissimilar according to the similarity matrix (Table S4). Improvement of the Involvement Score calculation is a matter of future research. In this work we tried to compensate these weak points by multiple comparisons of PKA to different protein kinases and combining all Involvement Scores.

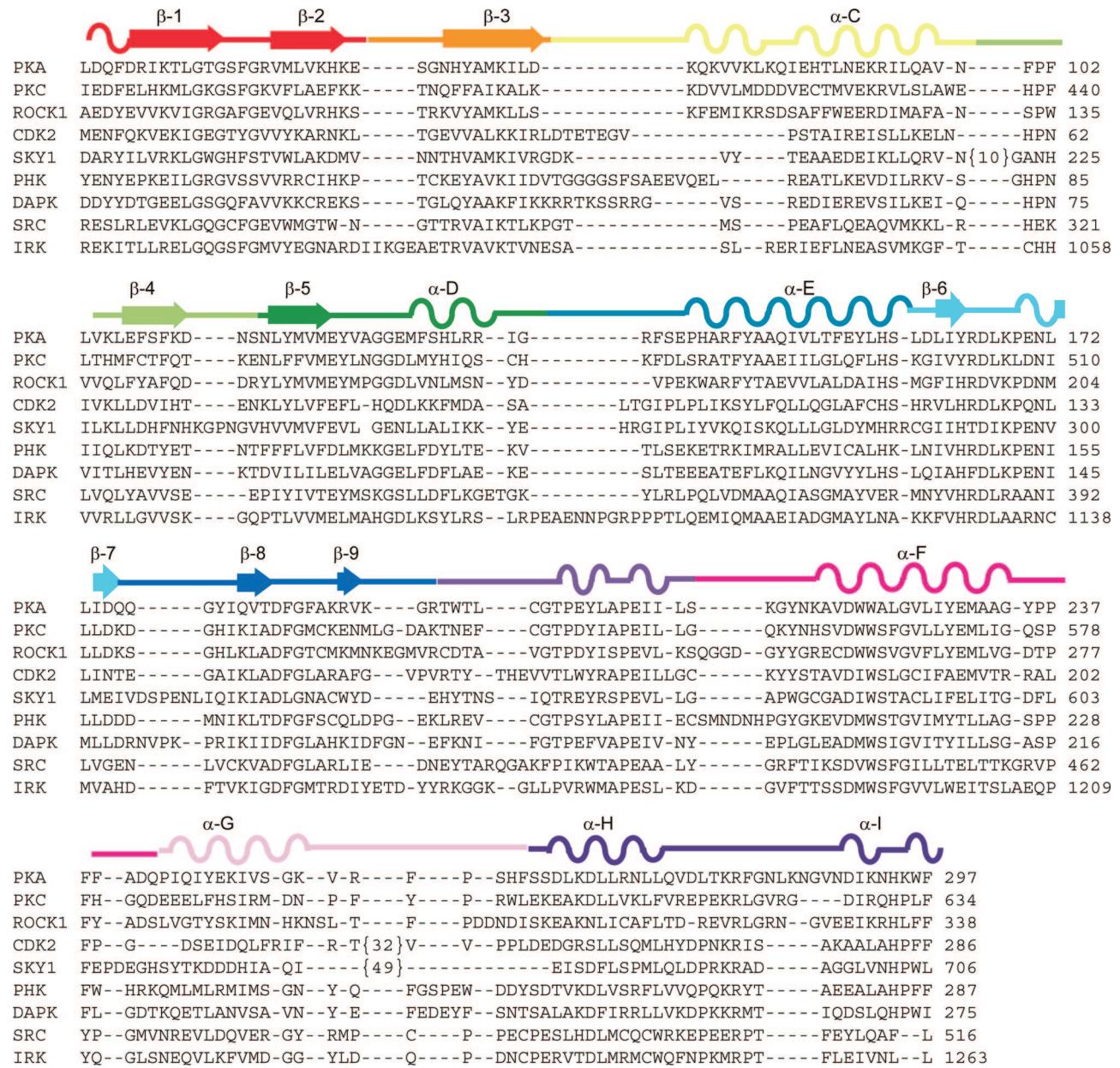


Fig. S1. Illustration of similarity between graphs A and B. Isomorphic subgraphs are colored red. Red vertices contain the same letters, red edges connect identical vertices.

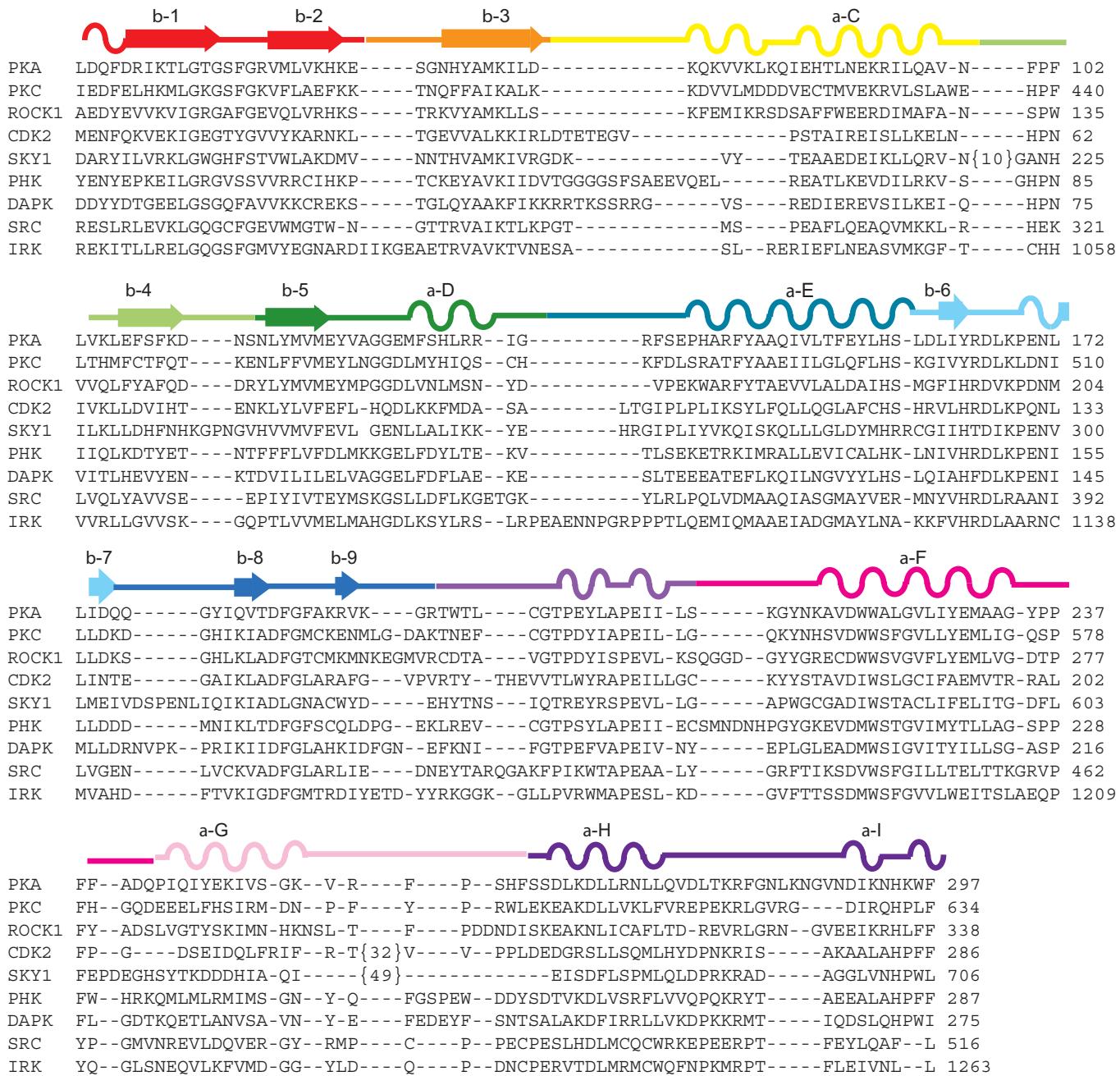


Fig. S2. Structure-based multiple sequence alignment of kinases used in the LSP-alignment. All protein kinases were aligned to PKA in pairwise way by Combinatorial Extension method [Shindyalov IN, Bourne PE (1998), *Protein Eng* 11:739–747]. The alignments were combined and manually curated to provide maximum consistency. Secondary structures of PKA are shown as colored cartoons. Different colors represent conserved subdomains of the kinase core.

Table S1. Involvement scores obtained by different residues of PKA in the LSP-alignment to eight different protein kinases

PKC	ROCK1	CDK2	SKY1	PHK	DAPK	SRC	IRK	Total
L-40			6					6
D-41	8	5	6			2		21
Q-42				10			2	12
F-43	13	11	15	6	11	11	2	69
D-44	12	14			11	13		50
R-45			11					11
I-46		14	8	9			10	41
K-47	10	16		12			9	47
T-48								0
L-49	13	9	11	14	17	19	9	104
G-50	9	11	8	12	8	12	9	83
T-51						10	2	12
G-52	3	6	3	8	7	8	5	40
S-53	2		5					7
F-54		4	3	6		2	6	23
G-55	7	10	5				7	37
R-56	18							18
V-57	26	29	15	18	11	23	11	146
M-58								0
L-59	16	31		14			12	73
V-60		19					2	21
K-61	19		12	12			3	60
H-62		8	12		10			30
K-63	3	2	5			2		12
E-64								0
S-65	4	9	9	5	4	5		36
G-66			7			6		13
N-67								0
H-68			12					12
Y-69	19	25			15	19		78
A-70	21	35	22	19	15	26	15	165
M-71	29	24	23	19			13	108
K-72	26	38	18	20	18	22	16	170
I-73		28		15	20			63
L-74	22	21	11	13	16	7	6	99
D-75					8			8
K-76	15	11				6		32
Q-77								0
K-78								0
V-79	15							15
V-80	8	9						17
K-81		6						6
L-82								0
K-83								0
Q-84								0
I-85	15							15
E-86	15			4	3	3		25
H-87								0
T-88	20				8			28
L-89	18		9		8		4	39
N-90				2			5	7
E-91	28	8	13	8	21	12	10	112
K-92	15	6						21
R-93	20			5			14	39
I-94	17	6	10	13	17	15	10	103
L-95	11	9	18	6	13	10	12	88
Q-96			11	4	12	8	4	47
A-97								0
V-98			4		11	11	9	35
N-99			6					6
F-100								0
P-101	20	4	7		11	3		45
F-102	28	3						31
L-103	31	24	22	13	22	23	26	172
V-104		27	11	9	19	18	12	107

	PKC	ROCK1	CDK2	SKY1	PHK	DAPK	SRC	IRK	Total
K-105		20	15	6	17		12	14	84
L-106	20	19	16	11	17	10	9	6	108
E-107									0
F-108		16							16
S-109	11	24			15				50
F-110	17	18		12		3			50
K-111	18	17							35
D-112		13				7			20
N-113		4			3				7
S-114					9	4			13
N-115	3			8		6			17
L-116	22	18	11	10	17	11	8		97
Y-117	27	28	16		16		16		103
M-118		34	14	21	15	22	19		125
V-119	24	32	22	23	20	18	16	12	167
M-120	22	29				13		16	80
E-121	19	23	19	9	12	18	19	14	133
Y-122	25	33	15				17		90
V-123	21		8	10		19			58
A-124						6	7	6	19
G-125	12	22		3		12			49
G-126	18	25			10	23	2	2	80
E-127	10	17	9		19	25		10	90
M-128	20	23	13	14	18	21	10	10	129
F-129				13	27	23	8		71
S-130		14		7				8	29
H-131	17				7			4	28
L-132	18	10	7	6	16	14	6	8	85
R-133	8			4			3	6	21
R-134									0
I-135	8							4	12
G-136									0
R-137	7								7
F-138	18				11	5			34
S-139					5	9		2	16
E-140		18				11			29
P-141			4				4		8
H-142				5					5
A-143	30	27				13	2		72
R-144		18	11	8	2				39
F-145	25	22							47
Y-146	37	21	12			23			93
A-147	34			19			7	7	67
A-148	35	23					12	9	79
Q-149	27	25	14	15		11	5	8	105
I-150	33	35	17	15	15	12	7	9	143
V-151	40	29	18	20	23	22			152
L-152	31	23		17					71
T-153									0
F-154	30	24	15	18					87
E-155	21	15		17					53
Y-156	24		17	16		12	11	14	94
L-157	22	19	22	16	20	16	15	16	146
H-158	20	17	17	13	10	17		12	106
S-159	13	13	7			10		7	50
L-160		12		7	12	10	7		48
D-161					13		6		19
L-162	19	6	13	12	18	12		9	89
I-163	39	22	17	23	25		23	20	169
Y-164	43	23	22	20	30	27	25	21	211
R-165	48	27	24		30		25	26	180
D-166	36	21	30	23	35	27	25	23	220
L-167	39	22	19	19	34	23	19	13	188
K-168	33	24	17	17	28	24	15		158
P-169		32	16	19	33	23			123
E-170	19	30	12	17	30	24		2	134

	PKC	ROCK1	CDK2	SKY1	PHK	DAPK	SRC	IRK	Total
N-171	27	20	25	18	36	23	17	14	180
L-172	25	34	24	22	28	25	16	13	187
L-173	33	32	10	15	24	27	12	9	162
I-174	29	17	5	8	15	23	9	10	116
D-175	22	11	3		18				54
Q-176	22	8					3		33
Q-177									0
G-178	17	10							27
Y-179	26	19							45
I-180	31	22	3	11	20		3	4	94
Q-181	24	14	9	10	24	6	11	2	100
V-182	22	24	20	14	29	24		13	146
T-183					31				31
D-184	26	30	23	21	45	32	21	16	214
F-185	24	14	21	20	29	26	24	13	171
G-186	18	12	24	17	32	27	21	14	165
F-187			21		28	22	17		88
A-188			20	12	21	17	20		90
K-189	7		18				14	10	49
R-190		4			17	7			28
V-191									0
K-192									0
G-193									0
R-194	10				6				16
T-195	17								17
W-196									0
T-197									0
L-198	17				14				31
C-199	16				7				23
G-200	22	19			25				66
T-201	20	16	3	9	24	7			79
P-202	20	19			21	9			69
E-203	23	21		15	29	19			107
Y-204	27	30	14	22	34	22	6	12	167
L-205	28	15			33	11	17	4	108
A-206	37	21	21	23	20	14	7	15	158
P-207	27	15	15	13	23	16	9	13	131
E-208	23	12	12	11	18	12		12	100
I-209	25	7	13	16	13	13			87
I-210	14	2	6	2		2		3	29
L-211	9		9	4			3		25
S-212									0
K-213	4								4
G-214									0
Y-215	25	8		3	17				53
N-216	24								24
K-217		16			19				35
A-218	29		23					17	69
V-219	37		21		29				87
D-220	39	27	26	21	29	20	21	17	200
W-221	38	29							67
W-222	33	28	26	23	28	26	13	19	196
A-223	43	40	16	21	37	24	23	19	223
L-224	46	44	31			28	18	20	187
G-225	38	41	25		30	20	14	21	189
V-226	38	39			35	26	18	19	175
L-227	35	42	19	16	33	27	17	19	208
I-228	32	33		19	27		15	17	143
Y-229	31	27		18	28	23		17	144
E-230	29	31	15	21			6	13	115
M-231	32	28	14	25	26	17	8	14	164
A-232									0
A-233					13	10		5	28
G-234	19	13		12	9	12			65
Y-235									0
P-236					24				24

	PKC	ROCK1	CDK2	SKY1	PHK	DAPK	SRC	IRK	Total
P-237	23	18			24	15			80
F-238	16	16	2	5	15	6			60
F-239		6			3	3			12
A-240		3							3
D-241									0
Q-242									0
P-243									0
I-244		7							7
Q-245	8					5			13
I-246	6				3		2		11
Y-247	12	9							21
E-248									0
K-249		11							11
I-250	22	21			11	8	4		62
V-251			2						6
S-252									0
G-253									0
K-254									0
V-255		9							9
R-256					4				4
F-257	6	6			5	8			25
P-258	5						2		7
S-259						6			6
H-260						4			4
F-261	3		2						5
S-262		8				10			18
S-263						8			8
D-264	13	8	11	5					37
L-265				7	12		10	6	35
K-266	18	12	14		11	8			63
D-267	8			2	12	8	4	4	38
L-268	20	10	15	6	11	15	7	9	93
L-269	27	13	15	14	10	12	10	8	109
R-270						12		10	22
N-271									0
L-272	25	16	22	11	16	17	10	10	127
L-273	19	24	19	12	22	25			121
Q-274			2	8			4	4	18
V-275				12	19				31
D-276	11		3	7		6		2	29
L-277									0
T-278									0
K-279	9		6	7	7	6			35
R-280	22	16	11	7	20	16	6	10	108
F-281	25	19			24				68
G-282	16	20							36
N-283									0
L-284									0
K-285									0
N-286									0
G-287									0
V-288		16							16
N-289									0
D-290		7			8	7		4	26
I-291	14	25					3	7	49
K-292	21	19					4		44
N-293				2				2	4
H-294	17	16	6	5	9	13			66
K-295									0
W-296		7	7	6	10	13			43
F-297	8	7	4	5	3				27

Table S2. Description of the 22 active protein kinases used for alignment of the C-spines in Fig. 2C

Name	Group	Require phosphorylation	RD kinase	PDB ID	Chain
Aurora A	AGC	Yes	Yes	1OL5	A
PKA	AGC	Yes	Yes	2CPK	E
PKB	AGC	Yes	Yes	1O6K	A
PDK1	AGC	Yes	Yes	1H1W	A
CDK2	CMGC	Yes	Yes	1FIN	A
ERK2	CMGC	Yes	Yes	2ERK	-
CDK5	CMGC	No	Yes	1H4L	A
CDK6	CMGC	No	Yes	1JOW	B
GSK-3 β	CMGC	No	Yes	1O9U	A
CK2	CMGC	No	Yes	1DAW	A
SKY1	CMGC	No	No	1HOW	A
IRK	TK	Yes	Yes	1IR3	A
IGF1RK	TK	Yes	Yes	1K3A	A
LCK	TK	Yes	Yes	3LCK	-
EGFR	TK	No	Yes	1M14	A
C-KIT	TK	No	Yes	1PKG	A
CSK	TK	No	Yes	1K9A	A
PHK	CAMK	No	Yes	2PHK	A
CHK1	CAMK	No	Yes	1IA8	A
DAPK	CAMK	No	No	1JKK	A
CK1	CK1	No	Yes	1CKJ	A
PKNB	Prokaryotic	Yes	Yes	1MRU	A

Table S3. Residues with high level of the Involvement Score (≥ 100) and their proposed functions (PKA sequence)

Residue	Total Involvement Score	Secondary structure	Function
L-49	104	β 1	ATP binding
V-75	146	β 2	C-spine
A-70	165	β 3	C-spine
M-71	108	β 3	Conserved hydrophobic feature of the N-lobe
K-72	170	β 3	ATP binding, α C-helix anchoring
E-91	112	α C	Anchors the α C-helix to K-72
I-94	103	α C	Anchors the α C-helix to the R-spine and β 6
L-103	172	α C- β 4 loop	Anchors the α C- β 4 loop to the R-spine
L-106	108	β 4	R-spine
M-118	125	β 5	Conserved hydrophobic feature of the N-lobe
V-119	167	β 5	Conserved hydrophobic feature of the N-lobe
E-121	133	β 5- α D loop	Conserved feature of the linker region
M-128	129	α D	C-spine
Q-149	105	α E	Anchoring β 8 to α E
I-150	143	α E	Anchoring β 8 to the F-helix
V-151	152	α E	Anchoring α E to α F and α I
L-157	146	α E	Anchors the E-helix to the R-spine
H-158	106	α E	Anchors the E-helix to the F-helix
I-163	169	β 6	Supports side chain of R-165
Y-164	211	Catalytic Loop	R-spine
R-165	180	Catalytic Loop	Activation Segment anchoring
D-166	220	Catalytic Loop	Catalysis
L-167	188	Catalytic Loop	Anchors the Catalytic Loop to the F-helix
K-168	158	Catalytic Loop	Catalysis
E-170	134	Catalytic Loop	Substrate binding
N-171	180	Catalytic Loop	Catalysis
L-172	187	β 7	C-spine
L-173	162	β 7	C-spine
I-174	116	β 7	C-spine
Q-181	100	β 8	? Possible anchoring to β 5- α D loop via E-121
V-182	146	β 8	Activation Segment anchoring
D-184	214	Activation Segment	Catalysis
F-185	171	Activation Segment	R-spine
G-186	165	Activation Segment	Positioning of D-184
Y-204	167	Activation Segment	Substrate binding, Activation Segment anchoring
L-205	108	Activation Segment	Substrate binding
A-206	158	Activation Segment	Activation Segment anchoring
P-207	131	Activation Segment	Activation Segment anchoring
E-208	100	Activation Segment	Activation Segment anchoring
D-220	200	α F	R-spine anchoring
W-222	196	α F	Activation Segment and α H-helix anchoring
A-223	223	α F	Catalytic Loop anchoring
L-224	187	α F	Catalytic Loop anchoring
G-225	189	α F	α H-helix anchoring
V-226	175	α F	Activation Segment and α H-helix anchoring
L-227	208	α F	C-spine
I-228	143	α F	α H-helix anchoring
Y-229	144	α F	α H-helix anchoring
E-230	115	α F	Substrate binding
M-231	164	α F	C-spine
L-269	109	α H	α H-helix anchoring
L-272	127	α H	α H-helix anchoring
L-273	121	α H	α H-helix anchoring
R-280	108	α H- α I loop	α H- α I loop anchoring

Residues with IS > 140 are shown in bold font.

Table S4. Amino acid substitution matrix used in the LSP alignment

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
R	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0
N	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0
D	0	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	0	0	0	1	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0
Q	0	1	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0
E	0	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
H	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0
I	0	0	0	1	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0	1
L	0	0	0	1	0	0	0	0	1	1	0	1	1	0	0	0	0	0	0	1
K	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0
M	0	0	0	1	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0
F	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	1	1	0	0
P	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
S	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0
T	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0
W	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	0
Y	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	1	1	0
V	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1

Residues were considered similar if they have score ≥ 12 in the optimized matrix presented by Hourai *et al.* [Hourai Y, Akutsu T, Akiyama Y (2004) Optimizing substitution matrices by separating score distributions. *Bioinformatics* 20:863–873]. We also took into account hydrophobicity of cysteine [Yang J, *et al.* (2005) Allosteric network of cAMP-dependent protein kinase revealed by mutation of Tyr204 in the P+1 loop. *J Mol Biol* 346:191–201.26] and considered it to be similar to methionine, leucine, and isoleucine.